Nucleic Acids Research, 2004, Vol. 32, Database issue ***D193–D195***
DOI: 10.1093/nar/gkh047

DomIns: a web resource for domain insertions in known protein structures

R. Aroul Selvam and Rajkumar Sasidharan[1,*]

The Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge CB10 1SA, UK and [1]MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, UK

Received August 4, 2003; Revised and Accepted September 22, 2003

ABSTRACT

Proteins can be formed by single or multiple domains. The process of recombination at the molecular level has generated a wide variety of multi-domain proteins with specific domain organization to cater to the functional requirements of an organism. The functional and structural costs of inserting a domain into another means that multi-domain proteins are usually formed by covalently linking the N-terminus of one domain to the C-terminus of the preceding domain. While this is true in a large proportion of multi-domain proteins, we find a significant fraction of proteins that are the result of domain insertion. The inserted domain breaks the sequence contiguity of the domain into which it is inserted leading to a novel domain organization. This web resource aims to document domain insertions in known protein structures that are classified in the SCOP database. The web server can be accessed from http://stash.mrc-lmb.cam.ac.uk/DomIns/.

INTRODUCTION

Domains constitute the basic structural, functional and evolutionary units of proteins (1–3). Proteins can be built from a single domain or from a combination of domains using a limited repertoire of domain families to form multi-domain proteins with widely diversified domain architecture and functions (4). Although most multi-domain proteins have a contiguous arrangement of domains, like beads on a string with one domain following the next in a sequential order, there are exceptions to this common pattern (5). Insertions are one form of non-contiguous domain organization, where one domain (insert) is inserted into another (parent) thus breaking the sequence contiguity of the parent domain. Although a few examples of domain insertions have been observed previously (6), the availability of an accurate and well curated domain classification in SCOP (7) and the exponential increase in the number of deposited structures in the PDB (8) now provides a platform to look into this intriguing structural organization of multi-domain proteins. We provide for the first time, a definite set of domain insertions in known protein structures through this resource.

IDENTIFYING INSERTIONS

In order to identify insertions, we used the SCOP (1) protein domain definitions and the SCOP parseable file, dir.cla.scop.txt_1.61, available here: http://scop.mrc-lmb.cam.ac.uk/scop/parse/index.html. Although there are several schemes for protein structure classification, we chose SCOP as it is a manually curated classification of proteins of known structures based on their structural and evolutionary relatedness. In SCOP, a protein domain is a unit of evolution if it occurs independently or in combination with other domains based on evidence from proteins of known structure. SCOP has a hierarchical classification scheme with the principal levels being family, superfamily, fold and class. Domains clustered together into families are clearly evolutionarily related, usually detectable at the sequence level. Families brought together into superfamilies may have low sequence identity, but their structural and functional features suggest that they have a common ancestry. Superfamilies with similar topology, but without evidence for evolutionary relatedness are grouped under a fold. Folds are classified into classes based on the secondary structure elements present. We only considered the major five classes (all-α, all-β, α/β, $\alpha+\beta$ and Small proteins), at the fold and superfamily levels of SCOP hierarchy in determining insertions. Considering only multi-domain proteins, we define a case of domain insertion if a domain occurs within a different domain of the same chain (Fig. 1). When more than one domain is inserted in a parent, we categorize them as multiple insertions. The domains involved in insertions can come from the same or different SCOP superfamilies.

ACCESS METHODS

The server can be accessed from the URL: http://stash.mrc-lmb.cam.ac.uk/DomIns/. There are various ways of retrieving information from the server.

(i) Search by PDB or SCOP identifier: a simple search identifies insertions given a PDB code with or without chain information or a valid SCOP domain identifier. No result for a given query may be due to the following reasons: (a) no known insertion, (b) there is no SCOP domain definition available for the PDB code or (c) the structure does not belong

*To whom correspondence should be addressed. Tel: +44 1223 402479; Fax: +44 1223 213556; Email: sraj@mrc-lmb.cam.ac.uk

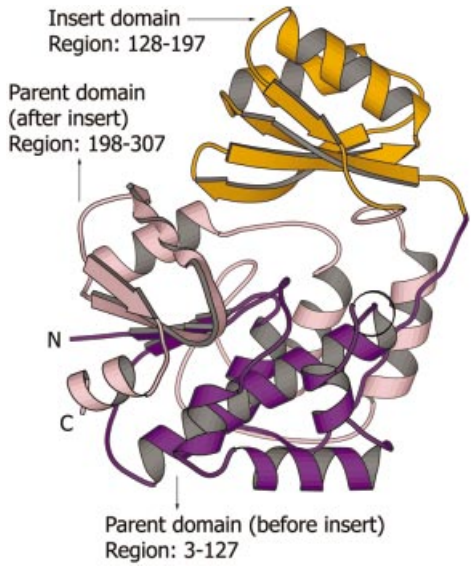

Figure 1. Domain insertion in malonyl-CoA: acyl carrier protein transacylase. The *Escherichia coli* protein malonyl-CoA: acyl carrier protein transacylase (PDB code: *1mla*) has two domains. The parent domain (coloured purple and pink) is interrupted by the insertion of the ACP-binding domain (insert—coloured orange). In the figure, the region of the parent domain (catalytic domain) preceding the insert domain (residues 128–197), with residues 3–127, is coloured purple and the region of the parent domain, which follows the insert domain with residues 198–307, is coloured pink. The parent and the insert domains belong to different SCOP superfamilies. A similar arrangement is also seen in *Streptomyces coelicolor* malonyl-CoA: ACP transacylase (PDB code: *1nm2*). This is an example of a single insertion, where the parent domain is interrupted by a single insert domain. In multiple insertions, there is more than one insert domain (e.g. *1zfjA*).

to any of the major five SCOP classes considered for identifying insertions.

(ii) Keyword search: this option allows the user to specify keywords (for example, D-amino acid oxidase) and retrieve a list of PDB entries with insertions that match the keyword(s).

(iii) Browsing all insertions: users can browse all known insertions one by one or choose to browse insertions from a non-redundant list of PDB chains. In order to have a representative sample of structures from PDB, we used a pre-computed list of non-redundant chains provided by PDB_Select (Apr 2002 release available from ftp://ftp.embl-heidelberg.de/pub/databases/protein_extras/pdb_select), with a sequence identity threshold of 90%. The procedure to extract such representative chains is explained in (9).

(iv) Search by insertion type: we categorized known insertions as single or multiple depending on the number of insert domains identified in a given chain. In single insertions, a domain belonging to a particular superfamily is inserted into another domain of the same or a different superfamily. In multiple insertions, more than one insert of the same or different superfamily is inserted into a parent domain. This search feature permits the user to display entries belonging to either or both of these categories.

(v) Search by SCOP class combination: we provide a search facility to retrieve entries with insertions based on the combination of SCOP classes. This facility will also retrieve all insertions for a given parent or insert SCOP class. Figure 2 provides a screen shot of results for the PDB code *1mla* (10).

We used MySQL and Java Server Pages (JSP) to create this resource.

SUMMARY OF KNOWN INSERTIONS

As of SCOP Release 1.61, there are 1332 PDB chains that have at least a single insertion, out of which 1143 chains have just a single insertion and 189 chains have more than one insertion. However, in the non-redundant list of chains, there is a total of 149 insertions, with 131 single insertions and 18 multiple insertions.

IMPLICATIONS AND FUTURE WORK

While artificial bifunctional and multifunctional proteins have been created by engineering end-to-end gene fusions, there are only a handful of examples where it has been possible to create multifunctional proteins by inserting whole domains into pre-existing ones (16). We believe that the list of naturally occurring domain insertions through this resource provides a valuable tool that can be used to undertake studies on the effect of domain insertions on protein folding and to expand the repertoire of multifunctional hybrid proteins.

In addition to providing regular updates in conjunction with SCOP releases, we intend to annotate known insertions at different levels. We are currently working on providing a graphical representation of the organization of the domains in a given chain and a wire plot of the PDB chain with information on secondary structure with unique colour coding for individual domains in order to provide a more detailed view of the structural features of insertions.

ACKNOWLEDGEMENTS

We thank Emma Hill and Madan Babu Mohan for several useful suggestions. R.A.S. and R.S. acknowledge financial support from Cambridge Commonwealth Trust and the Medical Research Council, UK.

REFERENCES

1. Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.
2. Holm,L. and Sander,C. (1996) Mapping the protein universe. *Science*, **273**, 595–603.
3. Orengo,C.A., Michie,A.D., Jones,S., Jones,D.T., Swindells,M.B. and Thornton,J.M. (1997) CATH—a hierarchic classification of protein domain structures. *Structure*, **5**, 1093–1108.
4. Chothia,C. (1992) Proteins. One thousand families for the molecular biologist. *Nature*, **357**, 543–544.
5. Wetlaufer,D.B. (1973) Nucleation, rapid folding and globular intrachain regions in proteins. *Proc. Natl Acad. Sci. USA*, **70**, 697–701.
6. Russell,R.B. (1994) Domain insertion. *Protein Eng.*, **7**, 1407–1410.
7. Lo Conte,L., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res.*, **30**, 264–267.
8. Berman,H.M., Battistuz,T., Bhat,T.N., Bluhm,W.F., Bourne,P.E., Burkhardt,K., Feng,Z., Gilliland,G.L., Iype,L., Jain,S. *et al.* (2002) The Protein Data Bank. *Acta Crystallogr. D*, **58**, 899–907.
9. Hobohm,U. and Sander,C. (1994) Enlarged representative set of protein structures. *Protein Sci.*, **3**, 522–524.
10. Serre,L., Verbree,E.C., Dauter,Z., Stuitje,A.R. and Derewenda,Z.S. (1995) The *Escherichia coli* malonyl-CoA:acyl carrier protein

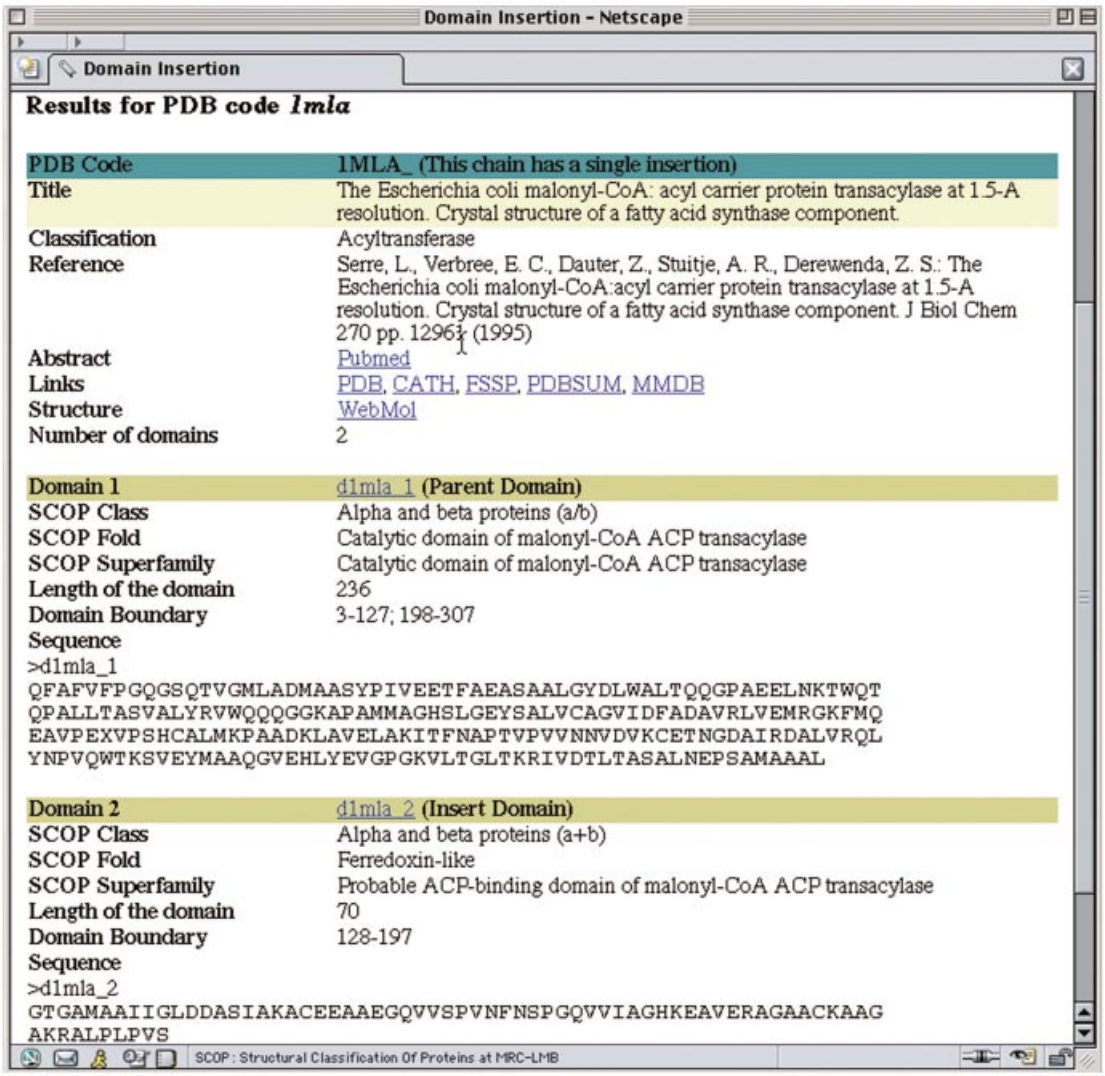

Figure 2. Screen shot of results for PDB code *1mla*. Each entry (PDB chain) in the database has the following information: the name of the protein, its biochemical function, the number of domains it contains, the SCOP classification for its individual domains, their sequence boundaries (based on SCOP domain definition), the Medline reference for the structure, sequence information, links to SCOP, CATH (11), FSSP (12), PDBSum (13), MMDB (14) and a Webmol (15) link to view the structure.

transacylase at 1.5-Å resolution. Crystal structure of a fatty acid synthase component. *J. Biol. Chem.*, **270**, 12961–12964.
11. Pearl,F.M., Bennett,C.F., Bray,J.E., Harrison,A.P., Martin,N., Shepherd,A., Sillitoe,I., Thornton,J. and Orengo,C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res.*, **31**, 452–455.
12. Holm,L. and Sander,C. (1998) Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.*, **26**, 316–319.
13. Laskowski,R.A. (2001) PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Res.*, **29**, 221–222.
14. Chen,J., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) MMDB: Entrez's 3D-structure database. *Nucleic Acids Res.*, **31**, 474–477.
15. Walther,D. (1997) WebMol—A Java based PDB viewer. *Trends Biochem. Sci.*, **22**, 274–275.
16. Betton,J.M., Jacob,J.P., Hofnung,M. and Broome-Smith,J.K. (1997) Creating a bifunctional protein by insertion of β-lactamase into the maltodextrin-binding protein. *Nat. Biotechnol.*, **15**, 1276–1279.